APPLICATION FOR A UNITED STATES PATENT

UNITED STATES PATENT AND TRADEMARK OFFICE

(MBHB CASE No. 00-1185; 3Com Case No. 3391.CS.US.P)

Title: **SYSTEM AND METHOD FOR TRAFFIC SHAPING BASED ON GENERALIZED CONGESTION AND FLOW CONTROL**

Inventors: Michael Freed, a citizen of Israel, and a resident of Pleasanton, California;

Satish Amara, a citizen of India, and a resident of Mt. Prospect, Illinois; and

Michael Borella, a citizen of the United States, and a resident of Naperville, Illinois.

Assignee: 3Com Corporation
5400 Bayfront Plaza
Santa Clara, CA 95052

Express Mail No.: EL604652587US
Date of Deposit: August 27, 2001



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

## FIELD OF THE INVENTION

The present invention relates to communications in computer networks. More specifically, it relates to a system and method for traffic shaping based on general congestion and flow control.

## BACKGROUND OF THE INVENTION

In high-speed networks, routers are required to meet certain quality-of-service requirements associated with communicating network entities. As is known in the art, the quality-of-service is a communication network attribute that exists between two network entities requiring a predetermined network service level. The basic network parameters typically affecting network performance are bandwidth and delays since the application traffic between two network entities in a computer network may require a certain minimum bandwidth or may be sensitive to delays. Thus, in the simplest terms, the quality-of-service means providing consistent and predictable data delivery services to communicating network entities requiring a predetermined level of network service. Further, the quality-of-service is the ability of a network element, such as an application process, host or router to have some level of assurance that its traffic and service requirements can be satisfied.

In general, the quality-of-service elements manage network resources according to application demands and network management settings and, thus, cannot provide certainty that resource sharing occurs. Therefore, quality-of-service with a guaranteed service level requires resource allocation to individual data streams through the network. In current implementations, a priority for quality-of-service developers has been to ensure that resources allocated to the best-effort traffic are not

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

limited after reservations are made. However, equally important is that the high-priority applications do not disable low-priority Internet applications.

The key mechanisms for providing a predetermined level of quality-of-service include an admission control, traffic shaping, packet classification, packet marking and packet scheduling. In quality-of-service enabled Internet Protocol ("IP") networks, it is necessary to specify a traffic profile for a connection or a set of connections. Traffic shaping or traffic conditioning is typically used, for example, to control the rate of data transmitted out of an interface so that it matches the speed of the remote target interface and, further, to ensure that the traffic conforms to a predetermined policy level. Thus, traffic shaping is primarily used to control the access to available bandwidth, to ensure that traffic conforms to a predetermined set of policies, and to regulate the flow of traffic in order to avoid congestion that can occur if the transmitted traffic exceeds the access speed of its remote, target interface.

Traffic shaping is typically implemented on an edge router or core router and provides a mechanism to control the amount and volume of data being sent into the network as well as the rate at which the data is being sent. The predominant methods for traffic shaping include a leaky bucket method and a token bucket method. The leaky bucket is typically used to control the rate at which data is sent into the network and provides a mechanism by which bursty data can be shaped into a steady data stream. The leaky bucket implementation is typically employed for shaping traffic into flows with a fixed rate of admission into the network and is generally ineffective in providing a mechanism for shaping traffic into flows with variable rates of admission.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

The token bucket provides a method for traffic shaping and ingress rate control. The token bucket provides a control mechanism that dictates when data can be transmitted based on the presence of tokens in a bucket and uses network resources by allowing flows to burst up to configurable burst threshold levels. In the token bucket implementation, tokens are "put" into the bucket at a certain rate, and the bucket has a predetermined capacity. In such an implementation, if the bucket fills up to its top capacity, newly arriving tokens are discarded. Similarly, if the bucket is full of tokens, incoming tokens overflow and are not available for future packets. Thus, at any time, the largest burst of data a source can send into a network is roughly proportional to the size of the bucket. In the token burst implementation, a system administrator may configure a token generation rate and a depth of the burst.

In addition to traffic shaping, the token bucket methods may be employed for congestion avoidance. As is known in the art, congestion avoidance refers to methods of controlling an average queue size on an outgoing interface of a router such as an edge router. The primary mechanism used by the token bucket and leaky bucket for shaping the traffic includes dropping the incoming data packets to the network. Some routers handle dropping the packets using a technique typically referred to as tail dropping. Using tail dropping, a router simply drops packets indiscriminately, i.e., without regard to priority or class of service, for example. Other methods that have been used to avoid congestion more effectively than tail dropping include a Random Early Detection ("RED"), a Flow-based Random Early Detection ("FRED"), or a Weighted Random Early Detection ("WRED").

When RED is not configured, output buffers fill during periods of congestion. When the buffers are full, tail drop occurs, and all additional packets are dropped.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

Since the packets are dropped all at once, global synchronization of Transmission Control Protocol hosts can occur as multiple hosts reduce their transmission rates. As the congestion clears, the Transmission Control Protocol sources increase their data transmission rates, resulting in waves of congestion followed by periods where the transmission link is not fully used.

The Random Early Detection mechanism controls the data congestion by dropping or marking packets with a drop probability. Typically, an algorithm used by the Random Early Detection mechanism may sample a queue length on a router and compare it to two threshold levels, a low threshold level and a high threshold level. For example, if the queue length is less than the low threshold level, no packets are dropped and packets are forwarded to a destination address. If the queue length is between the low threshold level and the high threshold level, incoming packets are dropped with a probability that is directly proportional to the queue length, and, if the queue length is greater than the high threshold level, all incoming packets are dropped.

The Random Early Detection mechanism reduces the chances of tail dropping by selectively dropping packets when, for example, an output interface begins to show signs of congestion. By dropping some packets early rather than waiting until the buffer is full, Random Early Detection avoids dropping large numbers of packets at once and minimizes the chances of global synchronization.

The Weighted Random Early Detection generally drops packets selectively based on IP precedence so that packets with a higher IP precedence are less likely to be dropped than packets with a lower precedence. In such an implementation, higher priority traffic is delivered with a higher probability than lower priority traffic. The



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

Weighted Random Early Detection is more useful in the core routers of a network, rather than at the edge routers that assign IP precedence to packets as they enter the network.

While the Random Early Detection mechanism and the variation thereof have been widely studied and employed in the existing computer networks, they suffer from a number of disadvantages. The Random Early Detection mechanism as well as the Weighted Random Early Detection mechanism signal the source about the congestion by dropping the packets and have the ability to control only a predetermined type of data, specifically, Transmission Control Protocol data.

Thus, the need remains for a system and method for traffic shaping in computer networks.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

## SUMMARY OF THE INVENTION

In accordance with embodiments of the present invention, some of the problems associated with traffic shaping are overcome.

An embodiment of a method for traffic shaping in a computer network, according to the present invention, involves receiving at least one data packet on a traffic manager from a user network entity and, responsive thereto, calculating at least one flow control parameter using the at least one data packet from the user network entity. Next, the at least one flow control parameter is compared to at least three threshold levels including a committed threshold level, a control threshold level and a peak threshold level. According to one embodiment of the present invention, the traffic manager includes a traffic shaper. In such an embodiment, the method includes calculating a data packet rate on the traffic shaper. If a value of the at least one flow control parameter, i.e. a data packet rate, falls between the committed threshold level and the control threshold level, the method includes applying a link layer control mechanism to control data flow from the user network entity. The link layer control mechanism may involve controlling a transmit slot allocation for transmission from the user network entity. In a data-over-cable system, the traffic manager may employ a bandwidth allocation mechanism, such as MAP, to control upstream bandwidth allocation for the user network entity.

Another embodiment of a method for traffic shaping according to the present invention involves, receiving at least one data packet on a traffic manager including a traffic conditioner from a user network entity and, responsive thereto, calculating at least one flow control parameter such as a queue size on an outgoing interface using the at least one data packet from the user network entity. Next, the queue size is

compared to at least three threshold levels including a committed threshold level, a control threshold level and a peak threshold level. If a value of the queue size falls between the committed threshold level and the control threshold level, the method includes applying a link layer control mechanism to control data flow from the user network entity. The link layer control mechanism may involve controlling a transmit slot allocation for transmission from the user network entity. In a data-over-cable system, the traffic manager may employ a bandwidth allocation mechanism, such as MAP, to control upstream bandwidth allocation for the user network entity.

Another embodiment of a method for traffic shaping in a data-over-cable system involves, receiving at least one data packet on a traffic manager from a cable modem via an upstream communication link and, responsive thereto, calculating at least one flow control parameter, such as a packet arrival rate or a queue size on an outgoing interface. Next, the method involves comparing a value of the at least one flow control parameter to at least three flow control threshold levels including a committed threshold level, a control threshold level and a peak threshold level. If the value of the at least one flow control parameter falls between the committed threshold level and the control threshold level, the method involves controlling a bandwidth allocation for upstream transmission from the cable modem.

An embodiment of a computer system, according to the present invention, includes an input interface arranged to receive at least one data packet from a network entity via an upstream communication link, an output interface arranged to send the at least one data packet to an external network, and a traffic manager connected to the input interface and the output interface. The traffic manager is arranged to calculate at least one flow control parameter using the at least one data packet received from

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

the network entity and compares it to a set of flow control thresholds including a committed threshold level, a control threshold level, and a peak threshold level. The traffic manager is further arranged to apply a flow control mechanism to control data flow from the network entity is a value of the at least one flow control parameter falls between the committed threshold level and the control threshold level.

These as well as other aspects and advantages of the present invention will become more apparent to those of ordinary skill in the art by reading the following detailed description, with reference to the accompanying drawings.

## BRIEF DESCRIPTION OF THE DRAWINGS

Exemplary embodiments of the present invention are described with reference to the following drawings, in which:

Figure 1 is a functional block diagram illustrating a system architecture suitable for application of the present invention;

Figure 2 is a functional block diagram illustrating a data-over-cable system for traffic shaping and congestion avoidance according to an embodiment of the present invention;

Figure 3 is a block diagram illustrating an exemplary implementation of threshold levels for traffic shaping and congestion avoidance according to an embodiment of the present invention; and

Figures 4A-4B are a flowchart illustrating a method for traffic shaping in a data-over-cable system according to an embodiment of the present invention.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

## DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Figure 1 is a functional block diagram illustrating an exemplary embodiment of a network architecture 100 suitable for application to the present invention for traffic shaping based on flow control and generalized congestion. Figure 1 shows the architecture 100 including a data network 118, such as a public Internet Protocol (IP) network to which a router 120 is linked via a network connection 116. A user network entity, such as a client device 102, is linked to the router 120 via a communication link 104. The communication link 104, as will be described in greater detail below, may include a cable connection, a wireless connection, a dial-up connection, or any other network connection utilizing a network protocol including link-layer mechanisms for controlling data transmission from the client device 102.

In the network architecture 100 illustrated in Figure 1, the router 120 may include an edge router. Edge routers typically assign IP priority to packets as the packets enter the network so that core routers on the network may use the priority parameters specified in the packets to determine how to forward the packets to the next hop on the network. The router 120 includes a traffic manager 106 for traffic shaping according to one embodiment of the present invention, in which, upon a receipt of at least one data packet from the client device 102, the traffic manager 120 calculates at least one flow control parameter and compares it to at least three threshold levels including a committed threshold level, a control threshold level and a peak threshold level. If the value of the calculated flow control parameter falls between the committed threshold level and the control threshold level, the traffic manager 120 applies a link-layer control mechanism to control data flow from the client device 102, as will be described in greater detail below.

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

The traffic manager 106 includes a traffic shaper 110 and a traffic conditioner 112. Figure 1, as well as subsequent figures, illustrates an embodiment in which the traffic shaper 110 and the traffic conditioner 112 are separate entities, where the traffic shaper 110 controls data flow one an incoming interface 108, and the traffic conditioner 112 controls data flows on an outgoing interface 114. However, it should be understood that in a preferred embodiment the router 120 may include a single entity including the functionality of both the traffic shaper 110 and the traffic conditioner 112.

The traffic shaper 110 calculates a flow control parameter such as a data arrival rate on the incoming interface 108. When the traffic shaper 110 calculates the data arrival rate, it compares the data arrival rate with a predetermined set of traffic shaping thresholds. If the flow control parameter calculated on the traffic shaper 110 includes a data arrival rate, the traffic shaper 110 compares it to at least three threshold levels including a committed rate threshold level, a control rate threshold level and a peak rate threshold level. In such an embodiment, if the data arrival rate falls between the committed threshold level and the control threshold level, the traffic manager 106 enables a flow control mechanism to control the rate of data being sent from the client device 102. The traffic manager 106 controls the data sending rate of the client device 102 using link layer specific methods. The link layer specific methods that may be employed on the traffic manager 106 for slowing down transmission rate from the client device 102 may depend on a type of the communication link 104 between the client device 102 and the router 120, an embodiment of which will be described in greater detail below in reference to a data-over-cable system.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

Referring back to Figure 1, the traffic manager 106 further includes the traffic conditioner 112 coupled to the outgoing interface 114. According to an exemplary embodiment of the present invention, the traffic conditioner 112 calculates a flow control parameter, i.e., a congestion avoidance parameter such as a number of packets queued on the outgoing interface 114 as means for the congestion avoidance control. The traffic conditioner 112 may control the congestion by calculating an average queue size on the outgoing interface 114. Upon calculating the average queue size, the traffic conditioner 112 may compare it to a predetermined set of queue size thresholds. Alternatively, the traffic conditioner 112 may compute a number of packets on the outgoing interface 114 and compare the computed number to a set of packet number thresholds. In either embodiment, the traffic conditioner 112 compares the computed average queue size or the number of packets to at least three threshold levels: a committed threshold level, a control threshold level and a peak threshold level.

In such an embodiment employed on the traffic shaper 112, if the average queue size or the number of packets on the outgoing interface 114 falls between the committed threshold level and the control threshold level, the traffic manager 106 enables a link-layer flow control mechanism to lower the average queue size or the number of packets on the outgoing interface 114. Similarly to the mechanism employed for the traffic shaper 110, the link-layer flow control mechanism may depend on the type of the communication link 104 between the client device 102 and the router 120.

According to an exemplary embodiment of the present invention, the router 120 includes typical routing elements associated with conventional routers. Referring

to Figure 1, the router 106 includes a memory unit 124 and a processor 122. The memory unit 124 may include a conventional routing table, threshold levels for the traffic shaper 110 and the traffic conditioner 112, as well as a set of instructions including logic that may be retrieved by the processor 122 to calculate flow control parameters such as a packet arrival rate on the incoming interface 108 or an average queue size on the outgoing interface 114. The set of instructions may further include logic for comparing the calculated flow control parameters to a set of thresholds levels stored in the memory unit 124 and, further, for applying link layer control mechanisms for slowing data transmission from the client device 102.

In one embodiment of the present invention, the memory unit 124 may store more than one set of data rate levels for the traffic shaper 110 and more than one set of packet number or queue size threshold levels for the traffic conditioner 112. In such an embodiment, for example, different sets of packet rate threshold levels or queue size threshold levels may be used for different client devices. For example, the memory unit 124 may include a table mapping a network address, such as a MAC address, of the client device 102 to a predetermined set of thresholds. The memory unit 124 may further include a set of instructions for assigning priorities to incoming packets based on, for example, a policy associated with different types of packets, or a service policy associated with the client device 102.

Figure 1 illustrates one exemplary architecture 100 suitable for application of the present invention, however, it should be understood that more or fewer network elements could also be used. Further, those skilled in the art will appreciate that the functional entities illustrated in Figure 1 may be implemented as discrete components or in conjunction with other components, in any suitable combination and

configuration. For example, the traffic manager 106 is not limited to including a combination of the traffic shaper 110 and the traffic conditioner 112 and, depending on the application, it could include solely the traffic shaper 110 or the traffic conditioner 112.

Figure 2 illustrates a functional block diagram illustrating another exemplary embodiment of a network architecture 200 suitable for application of the present invention in a data-over-cable network environment. The system architecture 200 illustrates a bi-directional cable network 206 supporting a downstream data flow, i.e., data flow from a headend to a subscriber 202 connected to a cable modem ("CM") 204, and an upstream data flow, i.e., data flow from the subscriber 202 to the headend. The cable-television network headend is a central location responsible for sending and receiving cable signals in downstream and upstream directions. However, the illustrated architecture is not limited to the bi-directional cable network 206 and may also include a uni-directional cable network, in which an upstream connection from the subscriber 202 to the headend 210 may include a dial-up connection via a telephone network such as a Public Switched Telephone Network ("PSTN"). In such an embodiment, the CM 204 may include an integral telephone modem for connecting to a PSTN that may provide a communication link between the CM 204 and the headend. When the upstream connection from the subscriber 202 is utilized via the PSTN, the upstream communication link terminates on a Telephone Resource Access Concentrator ("TRAC") that may be located at the headend or have routing associations with network entities at the headend. However, it should be understood that the uni-directional cable systems are not limited to employing the

PSTN as the upstream link, and the upstream connection may include a wireless connection, a satellite connection, or other types of connections.

The headend includes a cable modem termination system (CMTS) 220 connected to the cable television network 206 via an interface 208. Figure 2 illustrates one CMTS; however, the data-over-cable system 200 may include a plurality of cable modem termination systems. Further, the CMTS 220 and any other network entity described herein may be duplicated in a serial or parallel arrangement to provide back-up in case of failure.

The CMTS 220 may include a Total Control hub by 3Com Corporation of Santa Clara, California, with a cable modem termination unit. The Total Control hub is a chassis with multiple networking cards connected by a common bus. However, the CMTS 220 may use different network servers as well. The cable network 206 may include a cable television network such as one provided by Comcast Cable Communications, Inc., Cox Communications, or Time-Warner Cable, for instance.

The CM 204 may be connected to the cable network 206 with a downstream cable connection or an upstream cable connection, depending on whether it operates in a bi-directional cable system or an uni-directional cable system. The CM 204 may be provided by 3Com Corporation of Santa Clara, California, for instance; however, different cable modems provided by other producers may also be used. Figure 2 illustrates one CM 204 connected to the CMTS 220; however, typical data-over-cable systems may include tens or hundreds of CMs that may be connected to the CMTS 220. In addition, the CM 204 is connected to the subscriber entity 202 such as a customer premises equipment entity ("CPE") including a personal computer system, a VoIP device, or a telephone, for instance. The CM 204 may be connected to the CPE

202 via a cable modem-to-CPE interface ("CMCI"). Figure 2 illustrates one CPE 202; however, the CM 204 may be coupled to multiple CPE entities.

The CMTS 220 is connected to a data network 218 such as a public IP network via a CMTS-Network System Interface ("CMTS-NSI") 216. The CMTS-NSI 216 provides an interface for data packets transmitted via the data network 218 to the CPE 202 and for packets from the CPE 202 transmitted via the CMTS 220 to a network entity on an external network.

In the data-over-cable system 200, the CMTS 220 includes a traffic manager 222 arranged to calculate at least one flow control parameter that may be used for traffic shaping and congestion control according to an embodiment of the present invention. As illustrated in Figure 2, the traffic manager 222 includes a traffic shaper 212 and a traffic conditioner 214. The traffic shaper 212 calculates and monitors a packet arrival rate on the upstream connection via the interface 208 from the CM 204, and uses at least three threshold levels to determine a control mechanism that may be applied for traffic shaping. Similarly to the system architecture described in reference to Figure 1, the traffic shaper 212 compares the packet data arrival rate of packets at the interface 208 to at least three threshold levels such as a committed rate threshold level, a control rate threshold level, and a peak rate threshold level. If the calculated packet arrival rate falls between the committed rate threshold level and the control rate threshold level, the traffic shaper 212 enables a flow control mechanism to signal an end user, i.e., the CM 204.

The traffic conditioner 214 is coupled to the interface 216 and computes an average queue size or a number of packets on the interface 216. Similarly to the traffic shaper 212, the traffic conditioner 214 compares the computed number to at

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

least three threshold levels such as a committed queue size threshold level, a control queue size threshold level and a peak queue size threshold level. If the computed number falls between the committed threshold level and the control threshold level, the traffic conditioner 214 enables a link-layer mechanism to lower the queue size on the interface 216.

The traffic shaper 212 and the traffic conditioner 214 as illustrated in Figure 2 are separate entities. However, those skilled in the art will appreciate that the traffic conditioner and shaper may be implemented on a single network entity arranged to calculate and monitor a packet arrival rate from the CM 204 and a queue size on the outgoing interface 216, and the functionality of either entity may be enabled or disabled based on the need of a particular application.

Network devices for exemplary embodiments of the present invention illustrated in Figure 2 may interact based on standards proposed by the Data-Over-Cable-Service-Interface- Specification ("DOCSIS") standards from the Multimedia Cable Network Systems ("MCNS"), the Institute of Electrical and Electronic Engineers ("IEEE"), International Telecommunications Union-Telecommunication Standardization Sector ("ITU"), Internet Engineering Task Force ("IETF"), and/or Wireless Application Protocol ("WAP") Forum. However, network devices based on other standards may also be used. DOCSIS standards can be found on the World Wide Web at the Universal Resource Locator ("URL") "www.cablemodem.com." IEEE standards can be found at the URL "www.ieee.org." The ITU, (formerly known as the CCITT) standards can be found at the URL "www.itu.ch." IETF standards can be found at the URL "www.ietf.org." The WAP standards can be found at the URL "www.wapforum.org." However, the present invention is not

limited to these standards, and any other presently existing or later developed standard may also be used. Further, the data-over-cable system 200 may be compliant with Packet Cable specifications. The Packet Cable standards may be found on the World Wide Web at the URL "www.packetcable.com."

Typically, computer networks are described using the Open System Interconnection ("OSI") model that is a standard description or reference model for how messages may be transmitted between any two points in a communication network. The OSI model divides the process of communication between two end points into layers, where each layer is associated with a set of predetermined functional operations. Specifically, the OSI model defines seven layers including, from lowest to highest, a physical layer, a link layer, a network layer, a transport layer, a session layer, a presentation layer, and an application layer. The physical layer conveys a bit stream through the network at the electrical and mechanical level, and provides hardware means for sending and receiving data on a carrier. The data link layer provides synchronization for the physical layer and furnishes transmission protocol knowledge and management. The network layer handles routing and forwarding of data between communicating network entities. The transport layer manages the end-to-end control or error-checking, and ensures complete data transfer. The session layer sets-up, coordinates, and terminates sessions between the network entities. The presentation layer converts incoming and outgoing data from one representation format to another. Finally, the application layer identifies communicating entities, quality of service, and performs user authentication.

In the bi-directional cable system 200 illustrated in Figure 2, the CM 204 is connected to the cable network 206 in a physical layer via a Radio Frequency ("RF")



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

interface. In one embodiment of the present invention, for a downstream transmission, the RF interface may have a channel bandwidth of about 6 to 8 Megahertz ("MHz"), for example. The link layer includes a Medium Access Control ("MAC") layer, a Point-to-Point ("PPP") layer and an optional link security layer. The MAC layer controls the access to a transmission via the physical layer. The PPP layer encapsulates network layer datagrams over a serial communication link. The network layer includes an Internet Protocol ("IP") layer and an Internet Control Message Protocol ("ICMP") layer. The IP is a routing protocol designed to route traffic within a network or between networks. The ICMP provides a plurality of functions such as error reporting, reachability testing, congestion control, and route-change notification, for example. The transport layer includes a User Datagram Protocl ("UDP") layer that provides a connectionless mode of communication with datagrams. The transport layer may also include other types of layers, such a Transmission Control Protocol ("TCP"), for instance.

Above the transport layer, there are: a Simple Network Management Protocol ("SNMP") layer, a Trivial File Transfer Protocol ("TFTP") layer, a Dynamic Host Configuration Protocol ("DHCP"), and an UDP manager. The SNMP is used to support network management functions. The TFTP layer is a file transfer protocol that is used to download files and configuration files. The DHCP layer is a protocol for passing configuration information to hosts, and the UDP manager distinguishes and routes packets to an appropriate service. However, more, fewer, or different protocol layers could also be used.

In the DOCSIS environment, the bandwidth for a downstream channel ranges from 6 to 8 MHz and is shared between CMs receiving data from a CMTS, an



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

upstream channel in a data-over-cable system is typically modeled as a stream of mini-slots, where a CMTS generates a time reference to identify slots and controls slot allocation for each CM in the system. In the exemplary system architecture 200 illustrated in Figure 2, the CMTS 200 may allocate the upstream bandwidth to the CM 204 using an allocation MAP mechanism. The allocation MAP is a MAC management message transmitted by the CMTS 220 on a downstream channel of the cable network 206, and describes the application of each upstream mini-slot. The MAP may define slots as reserved, contention or ranging. A reserved time slot is a time slot that is reserved for upstream data transmission for a particular CM, such as the CM 204. A contention time slot is typically shared between a number of CMs, and a ranging time slot is used for management purposes.

The CMTS 220 may transmit one or more MAPs to the CM 204 throughout the operational stage of the CM 204. When the CM 204 receives a MAP from the CMTS 220, the CM 204 transmits data from the CPE 202 based on the time slot definitions specified in the MAP. Further, using the MAP, a receiver on the CMTS 220 has the ability to predict when data transmissions will occur from the CM 204. In order for the MAP to be properly applied in the system, it is necessary that the CM 204 and the CMTS 220 are time-synchronized. In order to do that, the CMTS 220 sends a global timing reference and a timing offset to the CM 204. The CMTS 220 creates the global timing reference by transmitting a Time Synchronization ("SYNC") MAC management message at a nominal frequency of a downstream channel between the CMTS 220 and the CM 204. The SYNC MAC message includes a timestamp identifying the time of transmission from the CMTS 220. Upon the receipt of the message, the CM 204 compares the receipt time of the message with the timestamp in

the SYNC MAC and, based on the comparison, adjusts its local clock reference. During a ranging process, the CMTS 220 calculates the timing offset so that transmissions from the CM 204 are aligned to the correct mini-slot boundaries. In one embodiment of the present invention, the CMTS 220 and the CM 204 may exchange ranging request messages until the correct timing offset is set on the CM 204.

Figure 3 is a block diagram 300 illustrating an exemplary functional implementation of threshold levels according to the present invention. The block diagram illustrates a client device 302 communicating over a communication link 304 with a traffic manager entity 306 including the functionality of a traffic shaper and a traffic conditioner for controlling data transmission from the client device 302 to a data network 326 via a communication link 328.

The traffic manager 306 employs three threshold levels including a committed threshold level 308, a control threshold level 310, and a peak threshold level 312. The threshold levels may include transmission rate threshold levels for the use on the traffic shaper, and queue size threshold levels for the use on the traffic conditioner. Figure 3 illustrates one set of threshold levels; however, it should be understood that threshold levels for the use on the traffic shaper and traffic controller may be different. Further, more than one set of threshold levels may be employed for the packet arrival rates, for instance. In such an embodiment, the traffic shaper may select a predetermined set of packet arrival rate threshold levels based on a source network device transmitting data packets.

Figure 3 illustrates two axes including a packet arrival rate axis 314 and a queue size axis 316. According to an exemplary embodiment of the present

invention, the traffic shaper and the traffic conditioner determine a packet arrival rate and an average queue size, respectively. Subsequently, the calculated values are compared to three threshold levels associated with the computed value. Thus, for example, when the traffic shaper computes a packet arrival rate using at least one incoming data packet, the traffic shaper compares the computed packet arrival rate with three packet rate threshold levels. If the computed packet arrival rate is below the committed threshold level 308, the functionality of the traffic shaper falls into a disable flow control region 318, and no action is taken on the traffic shaper. If the packet arrival rate is greater than the peak threshold level 312, the functionality of the traffic shaper falls into a packet drop region 324, and the traffic shaper drops the packets. Alternatively, the traffic shaper may be arranged to drop packets if the packet rate is greater than or equal to the peak threshold level. Further, if the computed packet arrival rate falls between the peak threshold level 312 and the control threshold level 310, the functionality of the traffic shaper falls into a packet drop with probability region 322. If the functionality of the traffic shaper falls into the packet drop probability region 322, the traffic shaper drops packets with a probability. In one embodiment of the present invention, the traffic shaper may determine that the packets should be dropped with probability if the computed packet arrival rate is greater than or equal to the control threshold level 310 and less than the peak threshold level 312.

The packet drop probability on the traffic shaper may be a function of the packet arrival rate, and may be determined using the following formula or any similar formula:

P = MaxP * Current_Arrival_Rate / Peak_Arrival_Rate;



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

In the formula, P is a drop probability, MaxP is a maximum drop probability, the Current_Arrival_Rate is a current arrival rate of packets of the incoming interface, and the Peak_Arrival_Rate is a maximum packet arrival rate. The MaxP parameter and the Peak_Arrival_rate parameter may be set by a system administrator, or, alternatively, may be pre-stored on the traffic manager 306.

Referring back to Figure 3, when the packet arrival rate falls between the committed threshold level 308 and the control threshold level 310, the functionality of the traffic shaper falls into a flow control region 320. In one embodiment of the present invention, the traffic shaper may determine that it should employ a flow control mechanism if the calculated packet arrival rate is greater than or equal to the committed threshold level 308 and less than the control threshold level 310. When the functionality of the traffic shaper falls into the flow control region 320, the traffic shaper employs link-layer methods to slow down the packet arrival rate from the client device 302.

The operation of threshold levels illustrated in Figure 3 has been described in reference to the traffic shaper and the packet arrival rates. However, it should be understood that the described embodiments would be equally applicable to the traffic conditioner utilizing queue size threshold levels instead of the packet arrival threshold levels.

Figures 4A and 4B illustrate an exemplary method 400 according to the present invention for employing the threshold levels illustrated in Figure 3 for traffic shaping in the data-over-cable system 200 of Figure 2.

Referring to Figure 4A, at step 402, a traffic shaper receives at least one data packet from a client device. In the embodiment illustrated in Figure 2, the traffic

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

shaper includes the traffic shaper 212, the client device includes the CM 204 communicating data from and to the CPE 202. Further, in the embodiment of Figure 2, the traffic shaper 212 receives the data from the CM 204 over an upstream communication link on the cable network 206.

Upon the receipt of data packets from the client device, at step 404, the traffic shaper computes a flow control parameter such as a packet arrival rate on an incoming interface, such as the interface 208 of Figure 2. At step 406, the traffic shaper applies at least three threshold levels to the computed packet arrival rate. In the embodiment illustrated in Figure 3, the traffic shaper compares the computed packet arrival rate to the committed threshold level 308, the control threshold level 310, and the peak threshold level 312. At step 408, the traffic shaper determines whether any action on the incoming packets is required. If the computed packet arrival rate does not exceed the committed threshold level 308, there is no action required, and, at step 410, the CMTS 200 forwards the received data packets to the output interface 216 and the data network 218.

If the packet arrival rate is greater than the committed threshold level 308, at step 412, the traffic shaper determines whether a flow control mechanism should be enabled. To do that, the traffic shaper determines whether the computed packet arrival rate is greater than or equal to the committed threshold level 308 and is below the control threshold level 310.

If the condition at step 412 is valid, at step 414, the CMTS enables a flow control mechanism such as a link-layer mechanism to control the packet arrival rate from the client device. In the data-over-cable system, the CMTS 220 controls the sending rate from the CM 204 by delaying or not generating reserved time slots for



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

data transmission from the CM 204. For example, when the CMTS 220 generates the next MAP for the CM 204, the generated MAP may not include any reserved time slots defining exclusive time intervals for upstream transmission from the CM 204. In such an embodiment, the CMTS 220 may allocate the time intervals that were not allocated to the CM 204 to another CM in the system.

Referring to Figure 4B, at step 416, the traffic shaper determines whether the client device has responded to the control mechanism. In a data-over-cable system, a CM uses reserved time slots for upstream transmission, however, if none of those are available, the CM may employ contention slots that are shared between all CMs utilizing the upstream channel. Thus, if the contention slots specified in the MAP are not employed by other CMs, the CM 204 may employ the available contention slots for the upstream transmission. Thus, to determine whether the control mechanism is successful, the traffic shaper may recalculate a packet arrival rate from the CM 204 and determine whether the computed rate decreased below the committed threshold level 308. If the packet data rate has decreased below the derived level, the method 400 terminates.

If the CM 204 fails to respond, at step 418, the traffic shaper determines whether a packet drop with probability should be enabled to slow down the packet arrival rate from the client device. To do that, the traffic shaper may use the recalculated packet arrival rate. If the recalculated packet arrival rate is greater than or equal to the control threshold level 310 and falls below the peak threshold level 312, at step 420, the traffic shaper computes a probability for dropping the incoming data packets from the CM 204. In one embodiment of the present invention the probability for dropping the packets may be calculated using the following equation:

P = MaxP * Current_Arrival_Rate / Peak_Arrival_Rate; where P is a drop probability, MaxP is a maximum drop probability set by a system administrator or prestored on the traffic shaper 212, Current_Arrival_Rate is an arrival rate of packets measured on the incoming interface, and Peak_Arrival_Rate is a maximum packet arrival rate set by a system administrator or prestored on the traffic shaper 212.

When the traffic shaper computes the packet drop probability, at step 422, the packet shaper starts dropping the incoming packets based on the calculated drop probability, and the method 400 may continue at step 402 in Figure 4A, where the traffic shaper continues monitoring the packet arrival rate from the CM 204.

Referring back to step 418, if the traffic shaper determines that the packet arrival rate is greater than or equal to the peak rate threshold level 312, the traffic shaper, at step 424, drops the data packets received from the CM 204, and the method 400 may continue at step 402, where the traffic shaper applies the method 400 to the next incoming packet.

The method 400 has been described in reference to the data-over-cable network 200 and the traffic shaper 212 on the CMTS 220. However, it should be understood that the analogous method could be applied on the traffic conditioner 214 arranged to monitor an average queue size on the output interface 216, and further to compare the average queue size to three threshold levels including a committed queue size threshold level, a control queue size threshold level, and a peak queue size threshold level. Further, based on the comparison, the traffic conditioner 214 may take an appropriate action, as described in reference to Figure 3. Thus, the traffic conditioner 214 may either forward the data packets to the public IP network 218, enable a flow control, drop packets with a predetermined probability or drop all

McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

incoming packets. If the average queue size is between the controlled queue size threshold level and the committed queue size threshold level and the method is applied in a data-over-cable system, the traffic conditioner 214 may decrease the number of contention slots momentarily for a data transmitting CM, or may identify the CM and then momentarily decrease the number of reserved slots for that CM.

It should be understood that the exemplary embodiments of the present invention are not limited to the data-over-cable systems and could be applied in other types of networks employing data link layers below the IP layer. Further, instead of employing the link-layer mechanisms on the traffic manager 106 or 122, the traffic managers may employ an Explicit Congestion Notification ("ECN") mechanism when a packet data rate or a queue size fall into the flow control region 320. The ECN mechanism provides a congestion indication mechanism, and a sender entity is notified of the congestion through a packet marking mechanism. To employ the ECN, an IP header of packets sent from the sender entity includes an ECN field having two or more bits. In such an embodiment, the sender entity may set an ECN-capable transport ("ECT") bit to indicate that the end points are ECN capable. If the traffic managers according to the exemplary embodiments detect that a packet data rate or queue size falls into the flow control region 320, the traffic manager may set a congestion experienced ("CE") bit to indicate congestion to the end nodes. When a receiving entity receives a data packet with a CE bit set, the receiving entity sends a congestion indication to the sending entity using, for example, an acknowledgement ("ACK") message having a CE bit set. When the sending entity receives a CE packet, the sending entity may reduce its congestion window. More information on the ECN mechanism may be found in a Request For Comments ("RFC") 2481, "A Protocol To



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001

Hold Explicit Congestion Notification (ECN) to IP", herein incorporated by reference and available from the Internet Engineering Task Force (IETF) at www.ietf.org.

Further, the embodiments of the present invention could be applied on a remote access server ("RAS") providing an Internet access to user network entities via dial-up connections. In such an embodiment, a traffic shaper and a traffic conditioner may be implemented on the RAS that may enable software flow control mechanisms to control data flow from the user network entities. Further, the embodiments of the present invention may be employed in Asynchronous Transfer Mode ("ATM") networks and Frame-Relay networks. In an ATM network, Explicit Forward Congestion Indication ("EFCI") bits may be employed to indicate a congestion state on a network entity, and in a Frame-Relay network Forward Explicit Congestion Notification ("FECN") bits may be employed to notify a network entity that, for example, a network router is experiencing congestion. For more information on ATM, see "B-ISDN ATM Adaptation Layer specification," ITU-I.363.3-1996, and "Functional characteristics of ATM equipment, ITU-I.732-1996, and "Functional Architecture of transport networks based on ATM," ITU-I.326-1995, incorporated herein by reference. Further, more information on the FECN mechanism may be found in the "ATM User Network Interface Specification V3.1," incorporated herein be reference and available from the ATM Forum and available at http://www.atmforum.com.

It should be understood that the programs, processes, methods and systems described herein are not related or limited to any particular type of computer or network system (hardware or software), unless indicated otherwise. Various types of

general purpose or specialized computer systems may be used with or perform operations in accordance with the teachings described herein.

In view of the wide variety of embodiments to which the principles of the present invention can be applied, it should be understood that the illustrated embodiments are exemplary only, and should not be taken as limiting the scope of the present invention. For example, the steps of the flow diagrams may be taken in sequences other than those described, and more or fewer elements may be used in the block diagrams. While various elements of the preferred embodiments have been described as being implemented in software, in other embodiments in hardware or firmware implementations may alternatively be used, and vice-versa.

It will be apparent to those of ordinary skill in the art that methods involved in the system for traffic shaping and congestion control may be embodied in a computer program product that includes a computer usable medium. For example, such as, a computer usable medium can include a readable memory device, such as a hard drive device, CD-ROM, a DVD-ROM, or a computer diskette, having computer readable program code segments stored thereon. The computer readable medium can also include a communications or transmission medium, such as, a bus or a communication link, either optical, wired or wireless having program code segments carried thereon as digital or analog data signals.

The claims should not be read as limited to the described order or elements unless stated to that effect. Therefore, all embodiments that come within the scope and spirit of the following claims and equivalents thereto are claimed as the invention.



McDONNELL BOEHNEN
HULBERT & BERGHOFF
300 SOUTH WACKER DRIVE
CHICAGO, ILLINOIS 60606
TELEPHONE (312) 913-0001